Another Point of View: Superiority, Noninferiority, and the Role of Active Comparators

Helena Chmura Kraemer, PhD

ABSTRACT

Despite substantial agreement with points made by Andrew C. Leon, PhD, in his article, I am not in complete agreement in a few areas. The definition of noninferiority proposed by Leon allows drugs somewhat less effective than placebo to be characterized as noninferior to placebo, and 2 active drugs may each be simultaneously noninferior to the other. Moreover, including a placebo arm in comparing 2 active drugs is of no use in deciding whether the study is well designed or not, since a significant difference between one of the active arms and the placebo may be due to chance or to a bias in the design. An alternative view of the situation is presented.


Corresponding author: Helena Chmura Kraemer, PhD, 1116 Forest Ave, Palo Alto, CA 94301 (hckhome@pacbell.net).

The topics addressed by Andrew C. Leon, PhD,1 are crucially important to clinical decision-making, not only in psychiatry but in all fields of medicine (particularly commercial drug development and US Food and Drug Administration decision-making).2 On the major issues, Leon and I strongly concur: the advocacy for comparative effectiveness clinical trials, the importance accorded active comparators so necessary to evidence-based medicine, and the emphasis on effect size (ES) rather than P values. However, there also remain some areas of disagreement and confusing issues.

Basic to the discussion is ES, a measure comparing the clinical impact on the patients in the population sampled, of the investigational drug (I) versus a control/comparison treatment (C), which may be placebo (P) or an active comparator (A). Leon uses standardized mean difference as the ES. Such a population ES is never known exactly, but it is estimated in randomized controlled trials (RCTs) with a certain margin of error conveyed by its confidence interval.

An ES would be zero if there were absolutely no difference between I and C, but such an occurrence is only theoretically possible, since by the time there are a theoretical rationale and an empirical justification for an ethical RCT comparing I versus C, there is little chance that the difference between them will be absolutely zero.3,4 However, the I versus C difference may well be too small to warrant any clinical preference for one intervention over the other, in which case I and C are clinically equivalent. Only if the ES is greater than some value d* is a strong clinical recommendation of one treatment over the other warranted. The top rows of Figure 1 represent traditional views of clinical superiority, inferiority, and equivalence; the bottom rows represent Leon’s understanding of inferiority and noninferiority.

The traditional valid 2-tailed hypothesis test at the 5% level of significance comparing I versus C typically requires that the chance of a statistically significant result, if ES were indeed 0, be less than 5%. An adequately powered valid hypothesis test at the 5% level of significance also requires that the chance of a statistically significant result be greater than, for example, 80%, whenever the ES is greater than d*. The parameter d* is typically called the critical value or the threshold of clinical significance.5,6 Leon calls his d* the threshold of clinical meaningfulness, which seems awkward, because +d* and −d* each lie within a noninferiority region (Figure 1).

Thus, on the traditional view, either I is preferable to C, or C is preferable to I, but not both. If I is preferable to C, I is either superior to or equivalent to C, and if C is preferable to I, C is either superior to or equivalent to I. In Leon’s view, if P is preferable to I, but the difference is not enough to be clinically significant (−d* < ES < 0), I is considered noninferior to P. If I were approved on the basis of its noninferiority to P, drugs less effective than placebo might be approved. Also, in the comparison of 2 active drugs, I and A, I might be found noninferior to A, and A noninferior to I, perhaps even in the same data set. This situation is bound to confuse.

However, the issue of whether to use the traditional 2-tailed hypothesis test at the 5% level of significance or the noninferiority test becomes irrelevant, if we agree that the goal of an RCT is not to show statistical significance but to estimate the ES comparing I and C.7
Leon is correct in pointing out how deficient we have been in setting the value of $d^*$, which depends on variables such as the seriousness of the indication, the danger of the consequences of inadequate treatment, and the costs and risks of the treatments. This deficiency has long been a problem, resulting in a proliferation of underpowered and often misleading RCTs. Generally, if $C$ is a placebo, $d^*$ would be set nearer .8, while if $C$ is an active comparator, $d^*$ would be set nearer .2. Since the sample size necessary for adequate power increases as $d^*$ decreases, the sample size for an $I$ versus $A$ comparison must typically be much larger than that for an $I$ (or $A$) versus $P$ difference. Thus, the adequacy of the design to detect an $I$ (or $A$) versus $P$ difference is no indication of the adequacy of the design to detect a difference in an $I$ versus $A$ comparison. Moreover, one cannot interpret finding any statistically significant result as proof of the adequacy of the design. Such a result may well arise by chance or because of design bias. The logic underlying the concept of *assay sensitivity* is flawed. But should one nevertheless include $P$ in any comparison of $I$ versus $A$ for another reason?

The ethical principle of *clinical equipoise* precludes proposing an RCT involving patients (a) in the absence of a theoretical rationale and empirical justification for the hypothesis to be tested or (b) after the answer is already scientifically known. The first criterion seems reasonably well-understood, but what does *scientifically known* mean? Surprisingly, researchers tend not to believe the results of certain comorbidities, at this stage those limits for if $I$ is not clinically superior to an intervention for if $I$ is not clinically superior to an intervention essentially does nothing ($P$) in the population in which $I$ is likely to have its *greatest effect*, it makes no sense to invest the time and resources to further develop it or to impose an unnecessary burden of multiple RCTs on patients.

Once it is known that $I$ is clinically superior to $P$ in the most favorable circumstances, effectiveness RCTs should follow, with a $P$, sampling the population with the targeted indication. The primary goal is to establish the ES in this target population, but of major concern also are moderators of treatment, *ie*, identification of subpopulations in which $I$ is clinically superior to $P$. It is quite possible that $I$ is clinically superior to $P$ in some subpopulation, and either equivalent to or less effective (harmful) than $P$ in others, which is the concern of personalized medicine. Thus, if $I$ is effective and safe only for those with a certain genotype, in a certain age range, in the absence of certain comorbidities, at this stage those limits should be established and clarified for medical consumers.
Once it is known that $I$ is clinically superior to $P$ in a specific subpopulation of those with the targeted indication, the question is whether there are other $A$s also known to be clinically superior to $P$ in that subpopulation. If so, one would seek to find in which subpopulations $I$ is clinically superior to $A$, in which subpopulations $A$ is clinically superior to $I$, and in which subpopulations $A$ and $I$ are clinically equivalent. (All 3 different subpopulations might exist.)

$A$ and $I$ should be dealt with symmetrically. One company’s $I$ is another company’s $A$, and neither should get preferential treatment. Until it is known that both $I$ and $A$ (active interventions) are clinically preferable to $P$ in the same subpopulation, it makes no sense to compare $I$ versus $A$ in an RCT in any population. As soon as it is known that both are clinically preferable to $P$ in a subpopulation, including a placebo control is unethical. Thus, $P$ should not be included in a study comparing 2 active treatments, $A$ and $I$.

Currently, the US Food and Drug Administration (and therefore drug companies) emphasizes $P$ values rather than ES, and it puts little emphasis on how representative the sample is of the population to which the results may be applied. There has been little attention to moderators of treatment response (the concern of personalized medicine). The emphasis is often on what can be shown in selected individual studies, not on the cumulative results of all valid RCTs done to date on a particular question. All of these approaches have a negative impact on the quality of medical decision-making, and they should be reconsidered.

Author affiliations: The Department of Psychiatry and Behavioral Sciences (Emerita), Stanford University School of Medicine; Palo Alto, California, and the Department of Psychiatry, University of Pittsburgh School of Medicine, Pennsylvania.

Potential conflicts of interest: None reported.

Funding/support: None reported.

REFERENCES


